APPENDIX 1

MEMORANDUM

TO: *Floyd* Plaintiffs Counsel
FROM: Jeffrey Fagan and Jack Glaser
DATE: November 20, 2018
RE: Remaining Concerns with the Monitor's Pilot Proposal

1. The Monitor's power analysis should account for the structure of the data that features non-independent observations, including both nested and group-based events.

   a) The Monitor team's power analysis may lead to a design that will generate unreliable estimates from a social science perspective. The power analysis did not account for the fact that the estimated 1,512 observations/police-citizen contacts are not independent of one another, or that they do not reflect the judgment of an individual officer under observation. Officers are likely to patrol in pairs or in teams, or an officer may be accompanied by a supervisor while on patrol. In either case, the observation of a stop event will reflect the shared judgment of two officers, or an interaction of the officers, or the influence of supervisors on officers. In each of these examples, it will be hard to attribute the observed officer's behavior solely to that individual, since others present during the patrol may influence the decision to make a stop, as well as the decision to characterize it as a Level 1, 2 or 3 stop.[1]

   b) A second concern is the fact that the unit of analysis is the officer, who will be observed across multiple stop events. In this way, observations are nested within officers and are not independent. The power analytic method that appears to have been employed assumes that the sample size represents independent observations (e.g., single behaviors by distinct individuals). But there will be multiple observations clustered within officer. Officers are routinely updating their stop knowledge, tying decisions in each event to the decisions and outcomes of recent prior events. Events, then, should be viewed as connected over time.

      Such clustering affects the Type I (false positive) error rate (alpha), which in turn affects the power estimates. Accordingly, the power analysis should take into account both the number of officers and the expected distribution of stops across officers.[2] Estimated differences in stop outcomes are likely to be biased when the study design fails to consider the cross-classified data structure of stops within officers.[3] While it might be argued that dependent measures research designs have greater statistical power, all else being equal, a valid power analysis for dependent measures requires the consideration of the actual sample size (how many officers). Because a given officer's behavior during

[1] For an illustration, see the Appendix in Jon B. Gould and Stephen D. Mastrofski, Suspect Searches: Assessing Police Behavior under the U.S. Constitution, *Criminology & Public Policy*, 3, 315-362 (2004) (describing the interactions among officers in making a suspect stop and search while on patrol).

[2] *See*, Aarts Emmeke, et al., A solution to dependency: using multilevel analysis to accommodate nested data. *Nature Neuroscience* 17, 491-496 (2014).

[3] E.M. Hoben et al., Measuring Disorder: Observer Bias in Systematic Social Observations at Streets and Neighborhoods, *Journal of Quantitative Criminology* 34,221-249 (2018). See, also, Albert J. Reiss, Jr., Systematic Social Observation of Natural Social Phenomena, *Sociological Methodology*, 3, 3-33 (1971).

one stop will be correlated with his or her behavior on both previous and then later stops, they cannot be treated as independent observations.

c) More generally, officer observations will be nested in groups (precincts, beats, or other administrative units). Failing to account for this lack of independence (and that resulting from repeated observations within officer) runs the risk of over-estimating the statistical power of the sample of 1,512 observed police-citizen contacts, since the number of independent observations will be modified by the extent of the joint participation of officers in civilian contacts.[4] This structural feature of the design of the pilot study suggests that additional power analyses be conducted to account for the lack of independence of observations.

2. To broaden the shared understanding of the power analysis, it would be helpful to convert the effect size results of the power analysis from the ".2 standard deviations" (Cohen's d) to a more meaningful and comprehensible statistic such as an odds ratio. A 0.2 Cohen's *d* is generally considered to be a "small" effect in the social sciences, so it is good that the power analysis has employed that threshold. However, it is difficult for the parties to translate that into an interpretable scale – what would it mean if officers in one experimental condition were 0.2 standard deviations less likely to conduct unconstitutional stops? An odds ratio can show how much more or less likely an officer is to conduct an unconstitutional stop or pattern of stops, or to fail to document stops in each of the 4 study conditions. Odds ratios provide a continuous measure of likelihood (as in, "group A is 1.5 times as likely as group B to…"), a more accessible measure that the binary measure based on a bright line of a standard deviation threshold. As a matter of social science, standard deviations provide less – or at least attenuated -- information about the differences between study conditions than do odds ratios. Given differences in sample distributions, the use of odds ratios provides a more accessible interpretation and comparison of potential results across conditions. Odds ratios can be provided alongside results based on SD's to more fully expand the basis of potential interpretation of results. With this information, stakeholders would be able to better assess the real "power" of the research design to detect effects of a magnitude that would matter.

3. The makeup of the "panel of experts" that will review and assess the legality of each police citizen investigative encounter should be diverse and independent from the operations of the monitor.

a) The analysis of the legality of police actions requires judgments that are framed by the institutional setting and specific practice domains of the individuals rendering judgments about the legal standing of a police action. Certainly, appellate courts do this by convening three-judge panels, where the judges themselves come from diverse practice backgrounds and jurisprudential perspectives. But more important, independent judgments often depart from the judgments rendered in a setting where there is deliberation among parties with different perspectives on law and evidence. There is ample evidence of the effects of deliberation on judgments, including in

[4] See, also, Steven W. Raudenbush, Statistical Analysis and Optimal Design for Cluster Randomized Trials, *Psychological Methods* 2, 173–185 (1997).

legal settings such as juries.[5] An optimal design for these panels would include participation of professionals with criminal defense and civil rights backgrounds, in addition to prosecutors, government lawyers, and current and former police officers.

b) Indeed, this has been the case and the norm in research on capital punishment. Recent empirical evidence submitted to courts on death eligibility has relied on panels of multiple raters who have reviewed evidence based on the facts of a case to determine whether the legal standard for death eligibility is met by the evidence.[6] By combining the expertise and unique perspectives of prosecutors, defense lawyers and judges, the ability to accurately classify a case as death eligible reflects the deliberations among professionals whose separate judgments may be biased or narrowed based on their institutional and professional experiences. A similar perspective can be applied in this case to de-bias judgments about the compliance of officers with legal standards for the conduct of investigative stops under state and federal caselaw.

c) The raters should also be independent of the Monitor team (as well as from the Floyd plaintiffs and NYPD). Members of the monitor team have been immersed not only in the logic and norms of police practice, but also in the natural tensions between the parties as well as between the parties and the monitor. In addition to their personal involvement, the members of the monitoring team bring their own perspectives on policy and practice, based on their institutional backgrounds and roles. These are strengths when translating research and evidence into policy and legal conclusions for the court, but they are potential sources of bias in a social science research project that requires neutral interpretation of complex evidence. Together with a deliberative process as described above, analysis of events by neutral raters are critical to ensuring the reliability of the results.

[5] *See, e.g.*, Dennis J. Devine, et al., Jury decision making: 45 years of empirical research on deliberating groups." *Psychology, Public Policy, and Law* 7, 622 - 727 (2001).

[6] See, Amended Declaration of David C. Baldus, *Troy Ashmus v. Robert K. Wong*, Case 93-CV-005694 (THE), November 18, 2010 (describing the use of multiple persons with legal training to code death eligibility from case documentation). See, also, *North Carolina v Marcus Robinson*, 91 CRS 23143 North Carolina Superior Court Division, 2012).